Survival Prediction of Titanic Passengers; Comparison of Random Forest and Decision Tree

Levent Ural
Computer Engineering Student
Bahcesehir University
Istanbul, Türkiye
levent.ural@bahcesehir.edu.tr
https://github.com/uralevent/ARI5001-Final-Project

***Abstract*—In this project, it is aimed to estimate the survival status of the passengers who traveled on the Titanic ship and were in the accident using the Titanic dataset. Random Forest and Decision Trees algorithms were used to make these estimates and the performance comparison of these two methods was made in the project. The missing data in the dataset was previously arranged and categorical values were converted to numerical values to facilitate the process. The performance comparisons of the two models used in the project were made with the accuracy rate, confusion matrix and classification report measurements. In addition, graphic visualizations were made in order to provide information about the dataset in the project. The aim of the project is to discover the positive and negative points of the algorithms used and to experience the applicability of the models in the real world.**

***Index Terms*—Depth-First Search (DFS), Breadth-First Search (BFS), A* Search, Heuristic Function, Manhattan Distance.**

I. Introduction

The Titanic disaster, which is also the subject of movies that most people know, is one of the greatest tragedies in the world. Unfortunately, more than 1,500 people lost their lives in this accident. Today, passenger demographics, ticket classes and survival probability data from this incident are an ideal data source for machine learning algorithms. These techniques work very well in finding patterns and similarities in such complex data.

Random Forest and Decision Trees algorithms, which will be used to analyze the data set and make predictions, will be compared in terms of their performance and prediction accuracy. Both algorithms are algorithms that successfully fulfill their duties. The prepared data set was made suitable by removing missing rows and converting variable values to more specific values before being included in the analysis.

The aim of the study is to evaluate algorithms such as Random Forest and Decision Trees in terms of their ability to predict survival on the Titanic. In addition, visualizations will be created in order to correctly evaluate the age, ticket class, port of embarkation and people in these categories in the data set.

II. DATASET

The dataset that will be used to test the performance of Random Forest and Decision Tree models during the project is the titanic.csv dataset. When the content of the dataset is examined, it includes information on 891 passengers. The categorized information of each passenger in the dataset is as follows;

- Survival (0 = No, 1 = Yes)
- Ticket Class (1 = 1st, 2 = 2nd, 3 = 3rd)
- Sex
- Age
- Ticket Number
- Passenger Fare
- Cabin Number
- Port of Embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)
- Sibling Ownership
- Spouse Ownership

During the evaluations, only Age, Sex, Ticket Class, Port of Embarkation, Survival and Passenger Fare were evaluated from the dataset. Cabin Number, Sibling and Spouse Ownership were not included since they would not be sufficiently effective in the evaluation.

III. METHODS

1) **Data Preprocessing**
In the first phase of the project, the data set was examined and attempts were made to eliminate missing parts. Failure to correctly eliminate missing parts in the data set may distort the results that the program will produce and lead to incorrect evaluations. The following procedures were performed to eliminate missing data;
   - **Missing items in the Age column:** The values in the Age column are filled with median values. The reason for using the median value instead of the mean value here is to reduce the effect of extreme values in groups where the median value may have extreme values such as age.
   - **Missing values in the Embarked column:** The values in the Embarked column are not numerical

values. The values in this column represent a category. For this reason, it would not be logical to use an average value. Instead, the most frequently occurring value should be placed here. The reason for this is that if the majority of the non-missing values are collected in one value, the majority of the missing values are also collected in this value, and the least erroneous arrangement can be made in this way.

- **Unresolved deficiencies:** There is a large amount of missing data in the Cabin number column. Since filling in this data could lead to erroneous results, this column was directly removed and excluded from the evaluation.

2) **Determining Independent and Dependent Variables**
While the dependent variable of the data set was set to survived, the independent variables were Ticket Class, Sex, Age, Fare and Embarked.
3) **Splitting the Dataset into Training and Testing**
In order to train the model and evaluate its performance, the dataset was divided into 80% training and 20% testing.
4) **Scaling of Data**
Having independent variables at different scales will affect the performance of the model. In order not to be affected by this situation, the StandardScalar function was used. In this way, standardization was achieved.
5) **Model Training and Evaluation**
The aim of the project is to compare Random Forest and Decision Trees algorithms. For this reason, these algorithms were first trained with the training set. After the training was completed, both models were tested using the previously divided test set. The models applied on this test set were evaluated using the following metrics:
   - Accuracy
   - Confusion Matrix
   - Precision, Recall, F1-Score
6) **Exploratory Data Analysis**
The connections and distributions of the passengers in the dataset with the variables are visualized. Survival rates according to age, gender and ticket class are analyzed and visualized.

IV. EXPERIMENTAL RESULTS

In this project, models were created by applying Random Forest and Decision Tree algorithms on the Titanic dataset. These models aimed to predict whether the passenger would survive by evaluating the passenger's age, gender, port of embarkation and ticket class. In this direction, when the models applied in this project were run, some results were obtained. The results and their comments are as follows;

1) **Data Analysis and Visualization**
   - **Age Distribution:** The age values of some passengers in the data are given as missing. In order to fill these values in a balanced and correct manner, the median values were placed in place of the missing data. The age distribution graph that was created as a result of this process is as follows. As can be seen here, the age distribution of the passengers on the ship is between 20-40. From here, we can see that giving the median value is a correct correction of the deficiency.

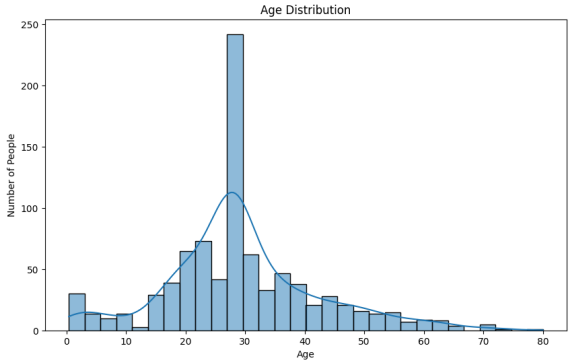

Fig. 1. Age Distribution

   - **Survival Rates by Gender:** As can be seen in the graph below, the survival rate of women is much higher than that of men. The reason for this is the procedure of rescuing women and children first in times of danger.

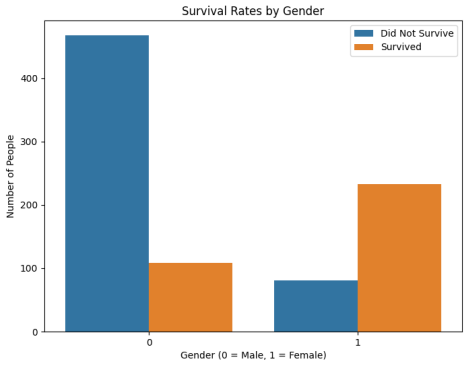

Fig. 2. Survival Rates by Gender

   - **Survival Rates by Ticket Class:** This graph shows the survival rates of passengers according to their ticket classes. Tickets are in 3 classes. First class tickets show the best class. As can be seen, the number of rescues decreases as the class decreases. The reasons for this may be that the areas where the classes are located are on higher floors or that higher classes have priority in rescue.

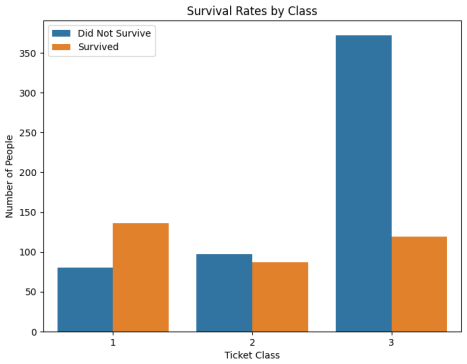

Fig. 3. Survival Rates by Class

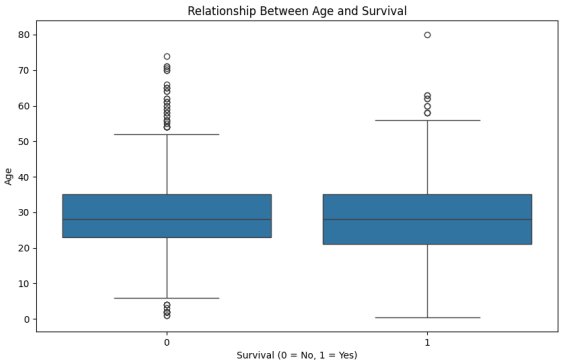

Fig. 5. Relationship Between Age and Survival

- **Survival Rates by Port of Embarkation:** Before the Titanic set sail, passengers were picked up at three ports. These were Cherbourg, Queenstown and Southampton. The economic conditions of the passengers who boarded at these ports may have determined the class of tickets they purchased. For example, Southampton is a place where more working class people lived in England. For this reason, lower class tickets may have been purchased. Since the higher class tickets were closer to the rescue boats, the lower class ones could not be saved. This makes it predictable whether the passengers who boarded at these ports would survive.

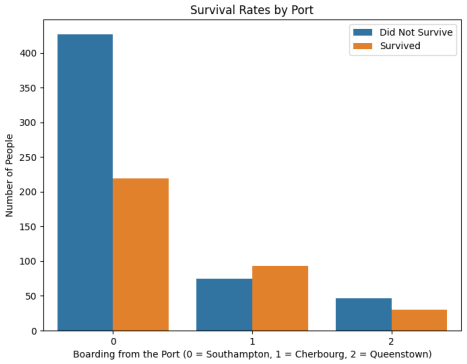

Fig. 4. Survival Rates by Port

- **Relationship Between Age and Survival:** When the survival rates are examined according to age, as can be seen in the graph below, the survival rates of older passengers are lower. Children survived the ship at a proportionally higher rate. This may be due to the fact that children, like women, are more likely to be rescued and loaded onto rescue boats due to their size.

2) **Model Performances**
The model performance results are as follows.

- **Decision Trees**
  - When the model created with the decision tree algorithm is examined, the accuracy value is 79.88
  - The Decision Trees model made 87 true negatives, 56 true positives, 18 false positives and 18 false negatives.

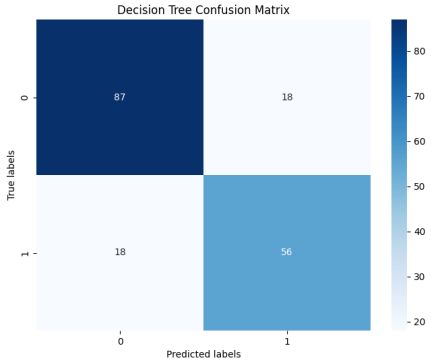

Fig. 6. Decision Tree Confusion Matrix

  - It performed well in metrics such as Precision, Recall and F1-score. The overall accuracy rate is around 80%.

```
Decision Tree Accuracy Score: 0.7988826815642458

Decision Tree Confusion Matrix:
 [[87 18]
 [18 56]]

Decision Tree Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.83      0.83       105
           1       0.76      0.76      0.76        74

    accuracy                           0.80       179
   macro avg       0.79      0.79      0.79       179
weighted avg       0.80      0.80      0.80       179
```

Fig. 7. Decision Tree Accuracy

- **Random Forest**
  - When the model created with the Random Forest algorithm is examined, the accuracy value is 82.12
  - The Random Forest model made 91 true negative, 56 true positive, 14 false positive and 18 false negative predictions.

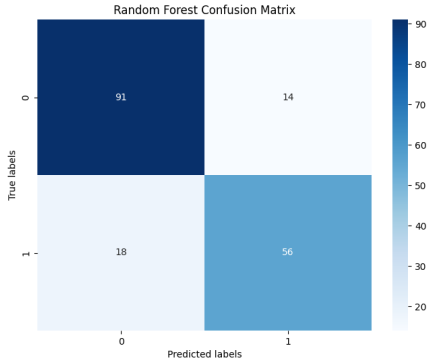

Fig. 8. Random Forest Confusion Matrix

  - It performed better than the Decision Trees algorithm in metrics such as Precision, Recall and F1-score.

```
Random Forest Accuracy Score: 0.8212290502793296

Random Forest Confusion Matrix:
 [[91 14]
 [18 56]]

Random Forest Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.87      0.85       105
           1       0.80      0.76      0.78        74

    accuracy                           0.82       179
   macro avg       0.82      0.81      0.81       179
weighted avg       0.82      0.82      0.82       179
```

Fig. 9. Random Forest Accuracy

3) **Comparative Assessment**
When the performance values of the models created and applied to the data set are examined, the Random Forest model performed better than the Decision Tree model as expected. The reason for this is that the Random Forest model is formed by combining the results of more than one Decision Tree model. Random Forest is proven to be a more stable model by the results. When the confusion matrices of the two models are examined, especially the false positive value shows a better result.

V. CONCLUSION

If a general evaluation is made, the Random Forest algorithm shows slightly better performance. Since this model is a combination of multiple Decision Tree algorithm results, it normally gave better results. Of course, the Random Forest model is a more complex model than the Decision Tree model. Considering the Titanic data set used, simpler models can also be used. However, in more complex data sets, Random Forest can be preferred because it will give more precise and clear results.

The created graphs and performance metrics also support these results. For better results, hyper parameters can be reviewed and edited.